CLAIMS

What is claimed is:

1. A method of determining content type of contents of a subject Web page, comprising the steps of:

providing a predefined set of potential content types;

for each potential content type, running tests having test results which enable quantitative evaluation of at least some contents of the subject Web page being of the potential content type;

mathematically combining the test results; and

based on the combined test results, assigning a respective probability, for each potential content type, that some contents of that type exists on the subject Web page.

2. A method as claimed in Claim 1 wherein the set of potential content types include any combination of organization description, organization history, organization mission, organization products/services, organization members, organization contact information, management team information, job opportunities, press releases, calendar of events/activities, biographical data, articles/news with information about people, articles/news with information about organizations and employee roster.

3. A method as claimed in Claim 1 wherein the step of combining includes producing a respective confidence level for each potential content type, that at least some content of the subject Web page is of the potential content type.

4. A method as claimed in Claim 1 wherein the step of combining the test results includes using a Bayesian network.

5. A method as claimed in Claim 4 further comprising the step of training the Bayesian network using a training set of Web pages with respective known content types such that statistics on the test results are collected on the training set of Web pages.

6. A method as claimed in Claim 1 wherein the predefined set includes a potential content type of press release; and

the step of running tests includes at least one of the following:

(i) determining whether a predefined piece of data or keyword appears in the subject Web page;

(ii) examining syntax or grammor or text properties;

(iii) examining page format and style;

(iv) examining links in the subject Web page; and

(v) examining the links that refer to the subject Web page.

7. A method as claimed in Claim 1 wherein the step of running tests includes any of:

(i) determining whether a predefined piece of data or keyword appears in the subject Web page;

(ii) examining syntax or grammor or text properties;

(iii) examining page format and style;

(iv) examining links in the subject Web page; and

(v) examining the links that refer to the subject Web page.

8. A method as claimed in Claim 1 further comprising the step of storing indications of the assigned probabilities of each potential content type per respective Web page.

9. A database formed by the method of Claim 8, said database containing indications of Web pages and corresponding content types determined to be found on respective Web pages.

10. Apparatus for determining content type of contents of a subject Web page, comprising:

a predefined set of potential content types; and

a test module utilizing the predefined set,

the test module employing a plurality of processor-executed tests having test results which enable, for each potential content type, quantitative evaluation of at least some contents of the subject Web page being of the potential content type, for each potential content type, the test module (i) running at least a subset of the tests, (ii) combining the test results and (iii) for each potential content type, assigning a respective probability that at least some contents of that type exists on the subject Web page being of the potential content type.

11. Apparatus as claimed in Claim 10 wherein the set of potential content types include any combination of contact information, press release, company description, employee list, other.

12. Apparatus as claimed in Claim 10 wherein the test module produces a respective confidence level for each potential content type, that at least some content of the subject Web page is of the potential content type.

13. Apparatus as claimed in Claim 10 wherein the test module combines the test results using a Bayesian network.

14. Apparatus as claimed in Claim 13 further comprising a training member for training the Bayesian network using a training set of Web pages with respective

known content types, such that statistics on the test results are collected on the training set of Web pages.

15. Apparatus as claimed in Claim 10 wherein the predefined set includes a potential content type of at least one of organization description, organization history, organization mission, organization products/services, organization members, organization contact information, management team information, job opportunities, press releases, calendar of events/activities, biographical data, articles/news with information about people, articles/news with information about organizations and employee roster.

16. Apparatus as claimed in Claim 15 wherein the processor-executed tests include at least one of:

(i) determining whether a predetermined piece of data appears in the subject Web page,

(ii) examining syntax or grammar or text properties,

(iii) examining page format and style,

(iv) examining links in the subject Web page, and

(v) examining links that refer to the subject Web page.

17. Apparatus as claimed in Claim 10 wherein the processor-executed tests include any of:

(i) determining whether a predetermined piece of data appears in the subject Web page,

(ii) examining syntax or grammar or text properties,

(iii) examining page format and style,

(iv) examining links in the subject Web page, and

(v) examining links that refer to the subject Web page.

18. Apparatus as claimed in Claim 10 further comprising storage means for receiving and storing indications of the assigned probabilities of each content type per Web page as determined by the test module, such that the storage means provides a cross reference between a Web page and respective content types of contents found on that Web page.